Journal of the Korean Data & Information Science Society
2014, **25**(5), 1095–1106

http://dx.doi.org/10.7465/jkdi.2014.25.5.1095
한국데이터정보과학회지

악성간암환자의 유전체자료 심볼릭 나무구조 모형연구[†]

이태림[1]

[1]한국방송통신대학교 정보통계학과

접수 2014년 7월 21일, 수정 2014년 8월 10일, 게재확정 2014년 9월 12일

요 약

본 연구에서는 악성간암환자의 생존기간에 영향을 주는 인자를 찾기 위하여 반응변수를 악성간암환자의 생존을 임상변수의 정보와 SNP유전인자를 통합한 자료를 대상으로 이해하기 쉬운 나무구조 생존모형과 심볼릭자료분석을 실시하여 영향을 주는 유의한 인자 뿐 아니라 그 임계치를 구하여 임상적으로 유용한 결과를 찾아 임상에 적용하는 것이 목적이다. 악성간암환자의 임상자료를 계량화하여 통계적 예후진단 모형을 구함으로써 임상변수 간 숨겨진 변수간의 관계를 규명하고 생존기간 군에 따른 예측 분류모형을 구하여 현시적으로 진단후 예후에 영향을 주는 중요 임상변수와 유전체변수 그 임계치를 구하여 임상에서의 치료계획에 중요한 근거를 제시했다. 심볼릭데이터 분석 결과 정상, 만성간염, 간염, 악성간염 등의 4개 군으로 구성된 1840명의 대상자를 분석 5 유전체의 20개 SNP가 밝혀진 바 있다. 즉 IL10-ht2가 악성간암의 발병에 매위 강한 관련이 있고 TGFB L10P-Prosms가 만성간염 환자 중 악성간암 발생 위험을 줄여주는 유전체로 밝혀졌다. SNP변수 와 질병군의 컴셉트 변수에 따라 상관정도를 원의 반지름 길이로 상대적으로 나타내 줌으로써 가장 판별력 있는 심볼릭변수를 상대적으로 비교할 수 있었다. 임상자료와 유전체자료를 통합하여 심볼릭 나무구조 생존모형을 구하여 생존기간을 군으로 한 나무구조모형을 유의한 변수와 기준치와 함께 구할 수 있었다.

주요용어: 나무구조모형, 심볼릭 자료분석, 악성간암, 유전체.

1. 서론

심볼릭 나무구조모형은 2004년도 발표된 Mballo와 Diday에 의한 심볼릭 의사결정나무모형의 시미노로브 기준치와 지니계수의 비교결과를 근거로 최적의 나무모형을 적합시키는 연구가 있었고, 일반적인 나무구조 생존분류모형은 악성간암 환자의 임상자료와 유전체 자료를 대상으로 생존유무 나무모형과 생존기간에 따른 분류모형 연구가 있었다 (Lee, 2005, 2012). 또한 골수암 환자의 5년 생존확률을 구하는 노모그램에 의한 암환자 생존기간 예측모형이 있었다 (Kim 등, 2009). 기존 연구에서 종속변인을 생존기간으로 하여 나무구조모형을 구한 바 있고 본 연구의 궁극적인 목표인 종속변수가 연속형 변수일 때의 심볼릭 생존나무모형을 적합시키기 위해서는 기존 연구에서 얻은 일반 생존나무모형에 의한 각 환자들의 소속 군을 종속변인으로 하여 이 소속 군을 근거로 심볼릭 의사결정 나무모형을 구하여 여기에 관여된 변인과 변인들의 조합인 컨셉트들의 특징과 악성간암환자 생존기간에 미치는 유의한 요인을 찾고자 한다. 2003년 Lynne과 Diday에 의해 발표된 SDA (symbolic data analysis)를 이용한 데이타마이닝의 이론을 기초로 심볼릭자료분석을 위해 본 연구 대상 자료인 간암 환자의 임상자료와 유전체자료를 가지고 생존나무 모형을 구축하고 이를 위한 변수의 정의와 변수의 조합을 구성하여 심볼릭 데이터 테이

[†] 이 논문은 2012년도 한국방송통신대학교의 연구비에 의하여 수행되었음.
[1] (110-791) 서울시 종로구 동숭동 86번지, 한국방송통신대학교 정보통계학과, 교수.
E-mail: trlee@knou.ac.kr

블을 구축하였다. 이 심볼릭 데이터 테이블을 근거로 하여 각종 기술적 정보와 2차와 3차 시각화에 의한 컨셉트 설명을 위한 변수 및 변수조합 들의 특징을 파악할 수 있었다. 본 논문은 심볼릭 자료분석의 이론과 나무구조 생존모형과의 적합타당성을 점검하고 심볼릭자료분석 이론과 유전체자료의 응용가능성을 타진하고 유전체 및 임상증상을 계량화한 정보를 기초로 한 심볼릭자료분석에 의한 간암이행예측 나무구조 분류모형, 생존모형을 구축하고 그 효율성을 통계적으로 검정하고자 한다. 즉 시간변화에 따른 예후인자 영향모형을 예측하고 다변량 요인들의 간암 전이기간 예측 통계모형을 탐색하고 심볼릭 나무구조 간암 이행모형을 적합하는 데 목적이 있다.

2. 연구방법 및 자료

2.1. 연구자료

분석대상인 HCC 임상자료 및 SNP자료는 2001년 1월부터 2001년 8월까지 서울대 내과 외래 및 간연구소에 등록된 B형 간염으로 인한 만성간염, 간경변, 간세포암환자 600명과 대조군 400명의 임상자료 및 혈액을 채취하여 산출한 SNP자료를 분석대상으로 했다. 임상자료 획득 및 혈액시료 채취를 위해 서울대학 병원을 내원한 환자를 대상으로 하여, HBsAg 양성환자는 최소 6개월마다 질환의 진행정도를 확인하였고, 크게 3 그룹 (만성간염, 간경화, 악성간암)으로 분류하여 그 중 악성간암 군을 모형적합의 대상으로 했다.

2.2. 변수

대상 환자들의 임상자료와 SNP 유전자료는 자료 수집 및 검토 대상의 변수들은 다음과 같은 절차로 수행되었다 (Lee, 2003). 임상변수를 성별, 연령 HBsAg, anti-HBs, anti-HBc, anti-HCV, HBeAg, anti-HBe, anti-HIV, AFP, 각종 초음파 및 CT, 혈관 조영 소견, 소변검사, 간기능 검사, 일반혈액검사 소견 등 임상변수 64종을 독립변수로 한다. 치료방법 별로 유의확률 0.05 기준으로 유의성을 검정하여 일차로 유의한 위험인자를 골랐고 임상자료의 전형적인 특징인 결측치가 많아 본 논문에서는 50%이상이 결측치인 경우는 변수를 제외하였다. 환자자료의 기본적인 임상변수의 특성은 다음의 Table 2.1과 같고 검토된 15개 후보 유전자의 다음 69개의 SNP에 대하여 스크린하여 한국인에게 우성으로 나타난 52개 SNP Table 2.2를 얻을 수 있었다.

Table 2.1 Pre-treatment baseline characteristics of HCC patient

Characteristics	Surgical Resection (N=91)	Transcatheter Arterial Chemoembolization (N=91)	P-value
Mean age ± SD	50 ± 10	56 ± 10	0.0002
Male sex - no. (%)	76 (84)	79 (87)	0.532
Serum alpha-fetoprotein - no. (%)			0.756
< 400 ng/ml	58 (64)	60 (66)	
400 ng/ml	33 (36)	31 (34)	
Viral marker - no. (%)			0.046
Hepatitis B virus	71 (78)	61 (67)	
Hepatitis C virus	4 (4)	16 (18)	
Hepatitis B and C virus	2 (2)	1 (1)	
NBNC&	13 (14)	10 (11)	
Unknown	1 (1)	3 (3)	
Okuda stage - no. (%)			0.155
Okuda I	91 (100)	88 (98)	
Okuda II	0 (0)	2 (2)	
UICC T stage - no. (%)			0.003
UICC T1	17 (19)	12 (13)	
UICC T2	63 (69)	49 (54)	
UICC T3	11 (12)	30 (33)	
CLIP scoring - no. (%)			0.125
CLIP1	54 (59)	39 (43)	
CLIP2	27 (30)	34 (37)	
CLIP3	8 (9)	16 (18)	
CLIP4	2 (2)	1 (1)	
CLIP5	0 (0)	1 (1)	
Lipiodol retention pattern - no. (%)			0.063
Compact	56 (62)	56 (62)	
Non-compact	30 (33)	35 (38)	
Unknown	5 (5)	0 (0)	
Serum alpha-fetoprotein in total alpha-fetoprotein secretion HCC patients	N=73	N=64	0.001
Decreased > 50%	61 (84)	33 (52)	
Decreased 25-50%	7 (10)	10 (16)	
Stable0	3 (4)	13 (20)	
Increased=25%	2 (3)	8 (13)	
Serum alpha-fetoprotein in alpha-fetoprotein secreting HCC patients with compact Lipiodol retention	N=41	N=40	0.137
Decreased > 50%	34 (83)	25 (63)	
Decreased 25-50%	4 (10)	6 (15)	
Stable	1 (2)	6 (15)	
Increased=25%	2 (5)	3 (8)	

Table 2.2 Screened polymorphisms and minor allele frequency

Gene	Loci	Genotype	p-value		
			allele freq.	heterozygosity	HWE
ACE	-5491 T >C	TT CT CC 425 557 203	0.406	0.482	0.676
	-5466 A >c	AA AC CC 430 562 202	0.405	0.482	0.729
	-93 T >C	TT CT CC 429 539 205	0.405	0.482	0.286
	-240 T >A	TT AT AA 444 568 211	0.405	0.482	0.450
	1237 T >C	TT CT CC 460 514 196	0.387	0.475	0.040
	2215 G >A	GG AG AA 507 623 204	0.386	0.474	0.858
	2350 A >G	AA AG GG 490 569 202	0.386	0.474	0.236
	3892 A >G	AA AG GG 412 538 220	0.418	0.487	0.172
	4656 Ims/del	ins ins/del del 436 569 236	0.419	0.487	0.119
NFKBIA	-673 T >C	TT AT AA 797 500 96	0.248	0.373	0.353
	- 642 C >T	CC CT TT 1081 226 67	0.131	0.228	0.000
	78 G >A	GG AG AA 848 483 85	0.231	0.355	0.347
	2444 G >A	GG AG AA 755 598 184	0.314	0.431	0.001
	2756 C >T	CC CT TT 640 574 168	0.329	0.442	0.086
	3053 C >T	CC CT TT 721 582 123	0.290	0.412	0.937
IL10	-1082 A >G	AA AG GG 1645 231 7	0.065	0.122	0.935
	-819 T >C	TT CT CC 734 595 118	0.287	0.409	0.986
	-592 A >C	AA AC CC 917 801 152	0.295	0.416	0.459
	+117 T >C	TT CT CC 1472 76 2	0.026	0.050	0.620
IL13	-1055 C >T	CC CT TT 1047 438 56	0.178	0.293	0.485
IL1A	-1203 C >T	CC CT TT 1083 190 11	0.083	0.151	0.709
	-899 C >T	CC CT TT 1281 219 11	0.080	0.47	0.888
	A135S G >T	GG GT TT 1324 228 9	0.079	0.145	0.971
IL1B	-511 T >	CC CT TT 341 628 355	0.505	0.500	0.176
	-31 C >T	CC CT TT 371 610 320	0.480	0.499	0.090
IL4	-1098 T >G	TT GT GG 1002 143 6	0.067	0.126	0.935
	-589 T >	TT CT CC 969 495 74	0.209	0.331	0.578
	-33 T >C	TT CT CC 975 490 64	0.202	0.323	0.970
IL5RA	-80 G >A	GG AG AA 830 566 163	0.286	0.408	0.000
IL6	-572 C >G	CC CG GG 833 539 99	0.251	0.376	0.650
TGFBI	-833 C >T	CC CT TT 1025 22 1	0.011	0.023	0.063
	-592 C >T	CC CT TT 395 752 352	0.486	0.500	0.987
	L10P T >C	TT CT CC 399 782 320	0.474	0.499	0.221
IL6RA	-183 G >A	AA AG GG 90 147 72	0.471	0.498	0.728
	24013 G >A	GG AG AA 242 62 6	0.122	0.214	0.768
	29753 A >C	AA AC CC 101 134 66	0.442	0.493	0.240
	42700 T >C	CT TT CC 139 106 66	0.383	0.472	0.000
	48869 T >A	TT AT AA 241 58 3	0.106	0.189	0.972
	48892 A >C	AA AC CC 104 135 65	0.436	0.492	0.239
	59818 C >T	CC CT TT 266 39 4	0.076	0.141	0.201
TNFA	-1031 T >C	TT CT CC 838 416 59	0.203	0.324	0.725
	-863 C >A	CC AC AA 1210 533 51	0.177	0.291	0.702
	-857 C >T	CC CT TT 1290 503 47	0.162	0.272	0.971
	-376 G >A	GG AG AA 1596 3 0	0.0001	0.002	0.999
	-308 G >A	GG AG AA 1710 164 4	0.046	0.087	0.999
	-238 G >A	GG AG AA 1690 173 19	0.056	0.106	0.000
	-163 G >A	GG AG AA 1557 28 16	0.019	0.037	0.000
TMFB	252 A >G	AA AG GG 455 644 302	0.445	0.494	0.034
	318 G >C	GG CG CC 648 503 228	0.348	0.454	0.000
HNF3	54 T >C	TT CT CC 1236 39 13	0.025	0.049	0.000
IFNG	-2671 T >C	TT CT CC 478 755 313	0.447	0.494	0.894
	-2459 A >G	AA AG GG 478 755 313	0.447	0.494	0.894

대부분의 임상자료의 특성과 같이 분석 대상 자료에서도 결측치가 많이 포함되어 있어서 결측치가 포함된 레코드를 모두 제거하면 너무 많은 정보가 손실되어 조사된 자료의 결측치를 포함 비율에 따라 적합시키고 그 정확도를 검토하여 자료의 손실을 최대한 줄이고자 시도했다. 적용된 결측치 추정 방법은 R-Cran에 탑재된 랜덤포레스트 추정법을 이용하여 결측치를 추정하여 분석하였다. 연속형 변수에 대해서는 가중치를 급접값으로 비결측치의 가중평균으로 구하고 범주형인 경우는 가장 큰 평균 근접값으로 보정하여 반복하여 결측치를 보정했다.

2.3. 나무구조 생존모형

나무구조 생존모형을 구하는데 이용된 STUDI는 survival tree of unbiased detection of interaction의 약자로 Loh와 Cho (2001)에 의해 개발된 나무구조 생존모형 알고리즘으로 종속변수가 악성간암 진단 후 생존기간으로 새로운 환자의 생존예측을 쉽게 할 수 있고 결측치에 대한 처리가 우수한 방법이다. STUDI는 생존함수를 다음과 같이 정의하여 이상치에 영향을 적게 받는 생존중앙치와 관찰치와의 차이에 의한 수정된 콕스-스넬 잔차를 구하여 유전요인의 유의성을 평가하는 척도로 한다. 수정된 콕스-스넬 잔차는 다음 식으로 정의된다.

$$S(y|X=x_i)=\exp\left\{-\Lambda_0(y)\exp(\beta^{'}x_i)\right\}$$

$$\widetilde{y}_i=inf\,\{y:S(y|X=x_i)\le 0.5\}:\text{생존시간}$$

$$R(t)=\sum_{i=1}^{n}|y_i-\widetilde{y}_i|:\text{노드 }t\text{에서의 리스크}$$

$$MCS=\widehat{\Lambda}_0(Y_i)\exp(\beta^{'}x_i)+0693(1-\delta_i),\ \ i=1,2,\ldots,n.$$

이 과정을 정리하면 다음과 같다.

① 각 변수를 각 노드에서 모형을 적합 시킨다.
② 각 노드에서 보정된 Cox-Snell 잔차을 구한다.
③ 변수군에 곡률검정 (curvature test)를 실시한다.
④ 쌍으로 된 n, s, f 변수군에 상호작용검정을 실시한다.
⑤ 가장 작은 값을 갖는 변수를 선택한다.

가장 작은 값을 갖는 변수 선택한다는 것은 곡률검정에서 가장 적은 값을 갖는 변수를 선택한다는 것이다. 상호작용검정에서 최소의 값을 갖는 변수를 선택하는데 이 경우 좀 더 작은 비용을 갖는 변수를 선정한다.

악성간암환자의 임상자료로 구한 나무구조 생존모형은 Figure 2.1와 같이 구하여 치료방법 즉 수술에 의한 방법과 TACE에 의한 치료방법에 따른 두 종류의 분류모형이 복합되어있는 것을 볼 수 있다 (Lee, 2012).

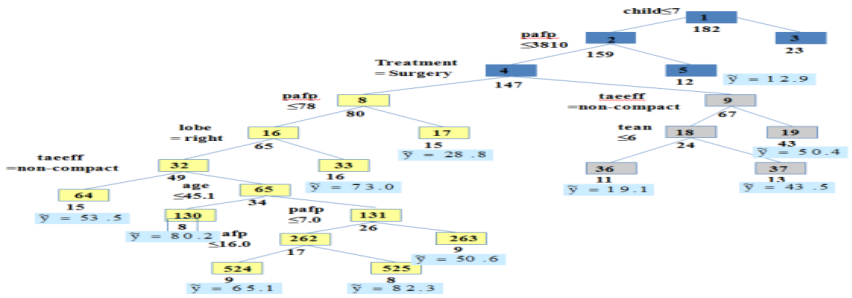


Figure 2.1 Tree structured survival model for HCC

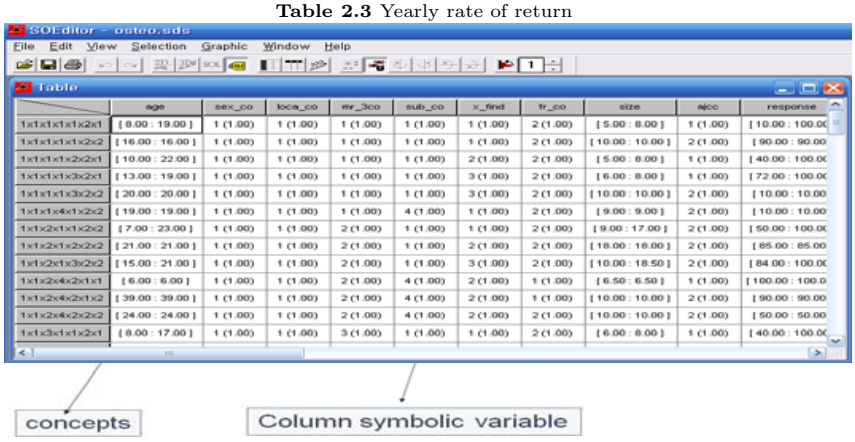
2.4. 심볼릭 자료분석 (SDA; symbolic data analysis)

데이터마이닝의 분석방법을 일반화 하여 심볼릭 데이터로 변수의 상위개념인 컨셉트 (concept)를 설명하고자 데이터베이스로부터 표준화 자료 표를 구축하여 새로운 정보를 찾고자한다. 여기서 상위개념을 컨셉트로 좀 더 복잡한 변수들 간의 관련성으로 설명하고자 한다. 즉 데이터마이닝으로부터 지식의 마이닝으로 발전 확대된 개념이다.

SDA의 단계는 다음과 같이 정리할 수 있다.
① 관련성있는 변수들을 정의하고 데이터베이스를 만든다.
② 데이터베이스 쿼리로부터 자료를 몇 개의 범주복합을 구성한다.
③ 여러 범주복합에 의해 개개 데이터를 컨셉트의 군으로 나눈다.
④ 심볼릭 자료표를 정의한다.
⑤ SDA 분석을 수행한다.

SDA를 위한 Syrokko 소프트웨어를 사용하기 위해 작성된 컨셉트와 심볼릭 자료 표는 다음과 같다.

Table 2.3 Yearly rate of return



심볼릭자료분석으로 구한 기술통계를 시각화한 그래프는 다음과 같다.

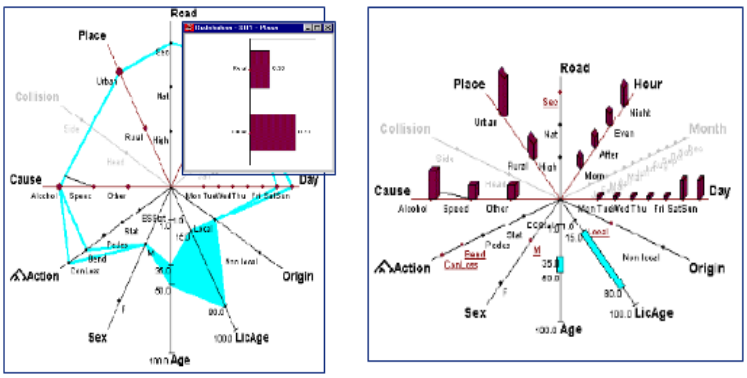


Figure 2.2 2D and 3D zoom stars graph

이러한 기술통계의 시각화를 통해 각 변수와 컨셉트의 특징을 파악할 수 있다. 연속형 변수의 신뢰구간은 축상의 구간으로 범주형 변수는 각 변수의 막대그래프로 그 특징을 현시적으로 파악할 수 있다. 심볼릭 자료분석으로 부터 얻은 심볼릭 트리모형은 다음과 같아 각 노드마다의 분기변수의 구체적인 정보를 파악할 수 있다.

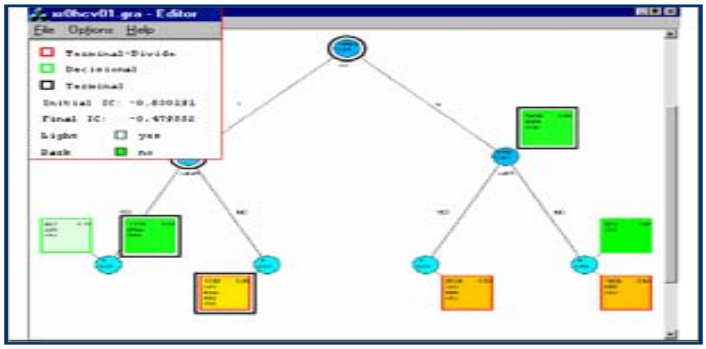


Figure 2.3 Symbolic classification tree

다음의 군집 나무모형에 의해 생존나무모형을 구할 수 있다.

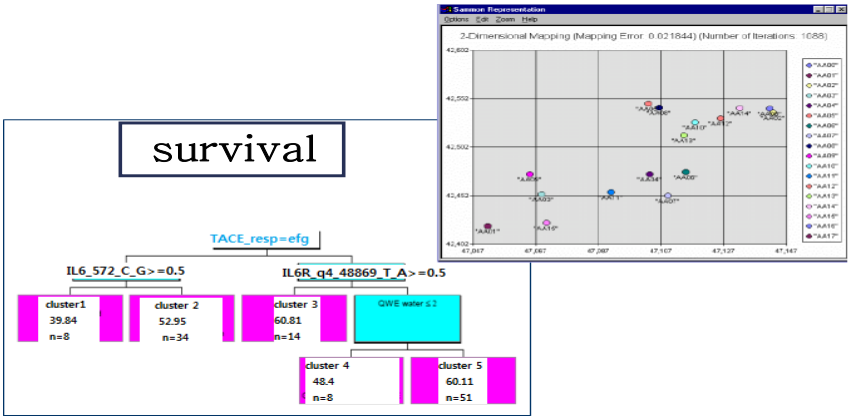


Figure 2.4 Symbolic survival tree

각 최종 노드마다 생존기간의 중앙치와 해당되는 표본의 크기를 알 수 있고 이를 근거로 최적의 집락수를 갖도록 군의 개수를 찾아 최적의 모형과 분류변수를 구할 수 있다.

3. 연구결과

3.1. 나무구조 생존모형

임상자료와 유전인자가 복합된 나무구조 생존모형을 근거로 악성간암환자의 생존기간에 영향을 주는 임상변수와 유전인자를 찾을 수 있었다. 또한 영향을 주는 변수의 임상기준치를 결과로 얻을 수 있고 대상 자료에 포함된 결측치의 추정과 포함 %에 따라 여러 유형의 생존 나무모형을 얻을 수 있었다.

가장 긴 생존기간을 보인 군은 67.18개월의 생존기간의 중앙치를 보였고 가장 짧은 수명기간은 44.41개월로 이 두 군으로 해당되는 데 관여하는 임상변수나 인자를 추적해본다면 환자의 생존기간을 연장하는데 기여하는 변수를 고려하여 치료계획을 세우는 데 근간이 될 수 있다.

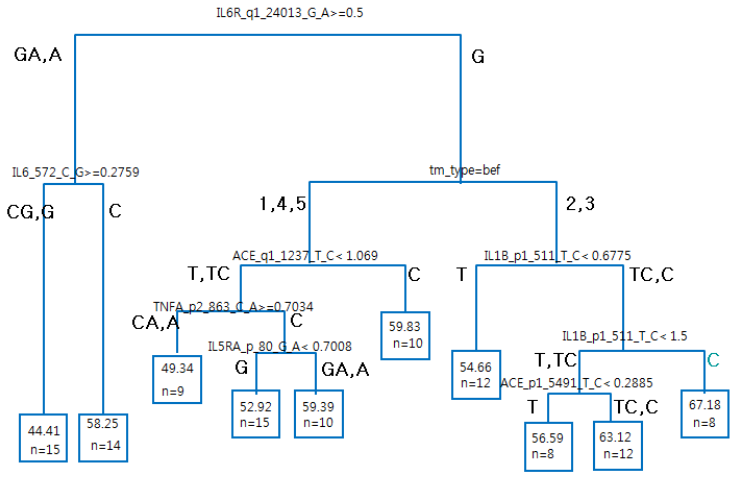


Figure 3.1 Tree structured survival model with SNP and clinical data with imputed 252 missing data

각종 결측치 보정에 따른 모형을 통해 정리된 유의 임상변수와 유전체변수의 요약표는 다음과 같다.

Table 3.1 Yearly rate of return

10%missing	10%imputation	20%missing	20%imputation	30%missing	30%imputation	fullmissing	full_imputation
IL6R_q1-24013	IL6R_q1-24013	IL1B_p1_511	IL1B_p1_511	IL6R_p1_183	IL6R_p1_183	TACE resp=ef	IL6R_p1_183
IL6_572	IL6_572	IL6R_q1_24013	IL6R_q1_24013	IL6R_q1-24013	IL6R_q1-24013	IL6_572	IL6R_q1-24013
IL1B_p1_511	tm_type=ef	TACE resp=ef	TACE resp=ef	IL6R_q1_24013	TGFB_q1_10	IL6R_q4_48869	TGFB_q1_10
tm_type=ef	ACE_q1_1237	IL6_572	IL6_572	IL6_572	IL6_572	IL6R_p1_183	IL10_p3_592
IL13_p_1055	IL1B_p1_511	IL13_p_1055	IL13_p_1055	TACE resp=ef	TACE resp=ef	IL6_572	IL6R_q2_29753
IL1B_p1_511	TNFA_p2_863	tm_type=ef	tm_type=ef	IFNG_q1_2459	IL1B_p1	ILIA_p2_889	PV_invasion
bid	IL5RA_p_80	IKB_p1_673	IKB_p1_673	TNFA_p2_863	IFNG_q1_2459	IL1B_p1_511	TACE resp=ef
	IL1B_p1_511	IL6R_p1_183	IL6R_q6_59818	TGFB_q1_10	TGFB_q1_10	TNFA_p2_863	TNFB_p1_318
	ACE_p1_5491	encephalohy	IKB_q4_3053	IL1B_p1_511	IL1B_p1_511	diureflc=b	IL6R_q6
			TNFB_p1_318	tm_type=ef		IL1B_p1_511	IL6_572
						IKB_p2	tm_type=ef
						TNM	IL6R_p1_183
						TNFA_p2_863	H6R_q4_48869
						tm_type=ef	TNFA_p1_1031
							IL1B_p1_511
							ACE_q1_1237

검토된 임상변수와 SNP변수 중에서 여러 모형에서 공통적으로 포함된 중요 분기변수 중 임상변수는 암의 유형인 암세포의 형태, TACE (내시경 치료) 횟수, 이뇨제 사용여부, encephalopathy (뇌질환) 등 이었고 그 외의 분기변수들은 SNP변수들로 IL6R_q1- 24013, IL6_572, IL1B_p1_511, IL13_p_1055, IKB_p1_673, IL6R_p1_183, TNFA _p2_863, IFNG_q1_2459 등이 유의한 변수로 골라졌다.

3.2. 심볼릭 자료분석

연속변수의 이산화 과정으로 분석결과 정의된 컨셉트에 의한 각 변수들의 분포가 Figure 3.2처럼 막대그래프로 나타내 군을 가장 잘 판별하는 변수의 분포를 시각적으로 보여주고 있다.

Figure 3.2 Characterization of the classes according to the evolution of HCC

악성간암에 걸리지 않고 간경화가 없으며 bCL군인 경우 (2*0*bCL) 진단명 3이고 간경화가 있고 bCL인 군 (3*1*bCL), 진단명이 4이고 간경화와 급성간암 (4*1* aHCC), 진단군 6와 간경화 증상이 없고 급성간암인 (6*0*aHCC) 군 등에 대한 각 변수들의 분포를 시각적으로 나타내주고 있는 것을 볼 수 있고 연령에 대해서는 95% 신뢰구간을 나타내주고 있다.

다음은 SDA PCA를 적용하여 요인축의 각 분면에 각 변수의 범주별 분포와 비중을 나타내어 비교할 수 있다. 각 군의 비중을 나타낸 다음의 Figure 3.3에서 HCC 진단명 3으로 간경화가 있고 bCL군은 비중이 낮은 지점에 위치하는 것을 상대적으로 평가할 수 있다.

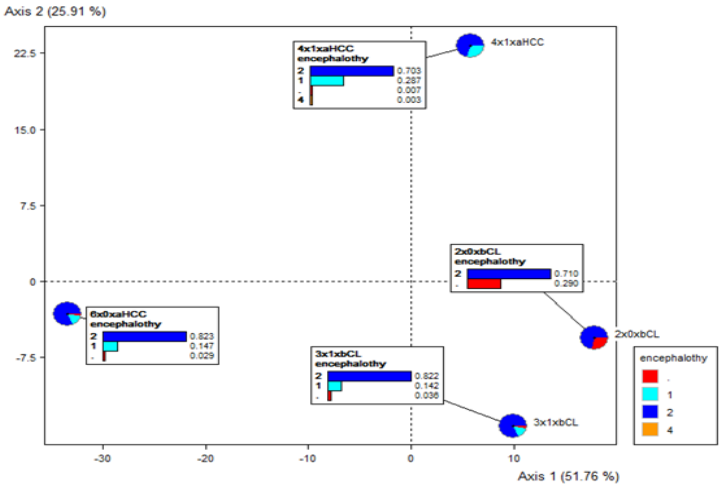


Figure 3.3 Maps showing the distribution in the categories of "Encephalothy", "Ascites", on the first factorial plane

Figure 3.4에서는 가장 판별력 있는 심볼릭 변수와의 상관관계를 원의 중심으로 부터의 축 거리로 나타낸 것으로 축의 길이와 방향으로 변수들의 사이의 상관성을 상대적으로 비교할 수 있다. 임상자료 HBeAg, encephalothy, 유전체 변수로 TGFBR3_126707_AG, CCND2_p171_TC 등이 상관성과 함께 판별력이 높은 변수로 상관원에 나타난 것을 볼 수 있다.

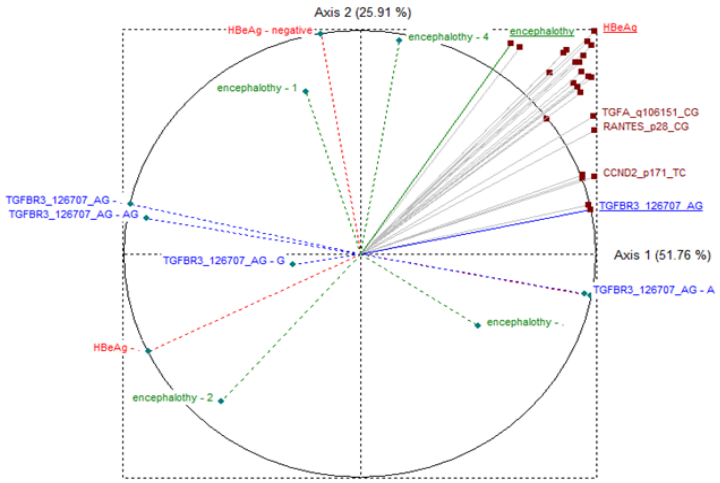


Figure 3.4 Correlation circle between the first more discriminating symbolic variables and some of their bins

Figure 3.5에서는 층화하지 않고 모든 임상변수와 유전체변수를 넣어 구한 상관도빈으로 변수들 간의 상대적인 상관도 비교를 중심으로부터의 거리축의 길이로 시각적으로 이해할 수 있다.

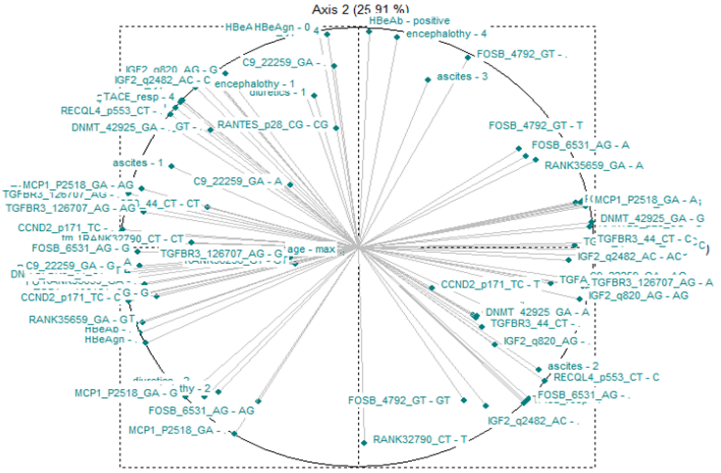


Figure 3.5 Correlation circle between the bins of the total symbolic variables

4. 결과 및 제언

본 연구에서는 악성간암환자의 생존기간에 영향을 주는 인자를 찾기 위하여 반응변수를 악성간암환자의 생존기간로 하고 임상변수의 정보와 SNP유전인자를 통합한 자료를 대상으로 나무구조 생존모형 및 심볼릭 자료분석 (SDA)을 실시하여 유의한 인자 뿐 아니라 그 임계치를 구하여 임상적으로 유용한 결과를 얻었다.

이전의 임상자료에 의해서만 구한 나무생존모형에 비해 복합적인 모형 결과를 얻을 수 있었고 변수의 중요도도 달리 평가되어 유전인자의 결합으로 인한 변수의 재평가를 할 수 있었다.

1. 악성간암환자의 임상자료를 계량화하여 통계적 예후진단 모형을 구함으로써 임상변수 간 숨겨진 변수간의 관계를 규명하였다.
2. 생존기간 군에 따른 예측 분류모형을 구하여 현시적으로 진단후 생존기간에 영향을 주는 중요 임상변수와 유전체변수 그 임계치를 구하여 임상에서의 치료계획에 중요한 근거를 제시했다.
3. 유전체 자료를 임상자료와 통합하여 진단 후 생존예측기간을 군으로 분류하는 생존나무구조모형을 구하여 임상자료 뿐 아니라 개개인의 유전자 자료에 의한 생존기간 특성을 규명했다.
4. 모형에서 중요 분기변수로 선택된 변수 중 임상변수로는 암세포의 형태, TACE (내시경 치료) 횟수, diuretics (이뇨제 사용)여부, encephalopathy (뇌질환)등 이었고 그 외의 유의한 유전체 분기변수들은 IL6R_q1-24013, IL6_572, IL1B_p1_511, IL13_p_1055, IKB_p1_673, IL6R_p1_183, TNFA_p2_863, IFNG_q1_2459 등이 유의한 변수로 골라졌다.

References

Afonso, F., Haddad, R., Toque, C., Eliezer E.-S. and Diday, E. (2013). *User manual of the SYR software*, Syrokko Internal Publication. Available from `http://www.syrokko.com`.

Breiman, L. (2003). *Manual for setting up, using, and understanding random forest V4.0*. Available from `http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf`.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistic of knowledge: Symbolic data analysis. *Journal of American Statistical Association*, **98**, 462.

Billard, L. and Diday, E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*, Wiley series in computational statistics, Wiley, Chichester.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic data analysis and the SODAS software*, Wiley, Chichester.

Diday, E. (2010). Principal component analysis for categorical histogram data: Some open directions of research. In *Classification and Multivariate Analysis for Complex Data Structures*, edited by B. Fichet, D. Piccolo, R. Verde and M. Vichi, Springer Verlag, New York.

Diday, E. (2012). Nonlinear canonical analysis for bar chart data tables and interpretation by coherency of meta bins and diversity of concepts. *Proceedings of 3rd Workshop in Symbolic Data Analysis*, 39-40.

He, Y. (2006). *Missing data imputation for tree-based models*, Ph. D. Thesis, University of California at Los Angeles, CA.

Kim, M. S., Lee, S. Y., Lee, T. R., Cho, W. H., Song, W. S., Cho, S. H., Lee, J. A., Yoo, J. Y., Jung, S. T. and Jeon, D. G. (2009). Prognostic effect of pathologic fracture in localized osteosarcoma: A cohort/case controlled study at a single institute. *Journal of Surgical Oncology*, **100**, 233-239.

Lee, H. (2003) *Searching for host genetic factors influencing the outcome of chronic HBV infection, especially the progression to hepato cellular carcinoma(HCC) by single nucleotide polymorphism (SNP) screening*, Project Report, 21C Frontier Research & Development Region, Seoul.

Lee, H. S., Kim, K. M., Yoon, J. H., Lee, T.R., Suh, K. S., Lee, K. U., Chung, J. W., Park, J. H. and Kim, C. Y. (2002). Theraputic efficacy of transcatheter aterial chemoembolization a compared with hepatic resection in hepatocellular carcinoma patients with compensated liver function in a hepatitisB virus-endemic area. *Journal of Clinical Oncology*, **20**, 4459-4465.

Lee, T. R. and Kim, M. J. and Myung, H. (2006). Independent prognostic factors of 861 cases of oral squamous cell carcinoma in Korean adults. *Oral Oncology*, **42**, 208-217.

Lee, T. R. and Moon, H. S. (1997). Classification of craniofacial patterns of children. *The Journal of Korea Society of Oral Health*, **21**, 54-65.

Lee, T. R. and Moon, H. S. (1998). Classification model for high risk dental caries with RBF neural networks. *The Journal of Data Science and Classification*, **2**, 38-47.

Lee, T. R. and Lee, H. S. (2009). Tree structured prognostic survival model for hepato cellular carcinoma using gene expression data. *Journal of the Korean Society of Health Information and Health Statistics*, **34**, 73-83.

Loh, W. Y. and Cho, H. (2006). Piecewise-constant tree-structured modeling for censored data. *Applied Statistics (Korea University Institute of Statistics)*, **21**, 31-53.

Mballo, C., Asseraf M. and Diday E. (2004). Binary tree for interval and taxonomic variables. *A Statistical Journal for Graduates Students*, **5**, 13-28.

Journal of the Korean Data & Information Science Society
2014, **25**(5), 1095–1106

http://dx.doi.org/10.7465/jkdi.2014.25.5.1095
한국데이터정보과학회지

Symbolic tree based model for HCC using SNP data[†]

Tae Rim Lee[1]

[1]Department of Information Statistics, Korea National Open University

Received 21 July 2014, revised 10 August 2014, accepted 12 September 2014

Abstract

Symbolic data analysis extends the data mining and exploratory data analysis to the knowledge mining, we can suggest the SDA tree model on clinical and genomic data with new knowledge mining SDA approach. Using SDA application for huge genomic SNP data, we can get the correlation the availability of understanding of hidden structure of HCC data could be proved. We can confirm validity of application of SDA to the tree structured progression model and to quantify the clinical lab data and SNP data for early diagnosis of HCC. Our proposed model constructs the representative model for HCC survival time and causal association with their SNP gene data. To fit the simple and easy interpretation tree structured survival model which could reduced from huge clinical and genomic data under the new statistical theory of knowledge mining with SDA.

Keywords: Hepato cellular carcinoma, SNP, symbolic data analysis, tree structured model.

[†] This research is partially supported by a KNOU research fund.
[1] Professor, Department of Information Statistics, Korea National Open University, Seoul 110-791, Korea. E-mail: trlee@knou.ac.kr